Errors-in-Variables Modeling of Personalized Treatment-Response Trajectories

Zhang, Guangyi

2021-01

Zhang , G , Ashrafi , R A , Juuti , A , Pietiläinen , K & Marttinen , P 2021 , '
Errors-in-Variables Modeling of Personalized Treatment-Response Trajectories ' , IEEE
Journal of Biomedical and Health Informatics , vol. 25 , no. 1 , pp. 201-208 . https://doi.org/10.1109/JBHI.2020.29873

http://hdl.handle.net/10138/331683
https://doi.org/10.1109/JBHI.2020.2987323

Errors-in-Variables Modeling of Personalized Treatment-Response Trajectories

Guangyi Zhang, Reza A. Ashrafi, Anne Juuti, Kirsi Pietiläinen, and Pekka Marttinen

***Abstract*—Estimating the impact of a treatment on a given response is needed in many biomedical applications. However, methodology is lacking for the case when the response is a continuous temporal curve, treatment covariates suffer extensively from measurement error, and even the exact timing of the treatments is unknown. We introduce a novel method for this challenging scenario. We model personalized treatment-response curves as a combination of parametric response functions, hierarchically sharing information across individuals, and a sparse Gaussian process for the baseline trend. Importantly, our model accounts for errors not only in treatment covariates, but also in treatment timings, a problem arising in practice for example when data on treatments are based on user self-reporting. We validate our model with simulated and real patient data, and show that in a challenging application of estimating the impact of diet on continuous blood glucose measurements, accounting for measurement error significantly improves estimation and prediction accuracy.**

***Index Terms*—Treatment-response trajectories, Bayesian methods, errors-in-variables, hierarchical models, Gaussian processes, wearable self-monitoring devices, time-series data.**

I. Introduction

INCREASING popularity of electronic health records (EHRs) and smart healthcare services has led to accumulation of large quantities of heterogeneous data, with potential to considerably improve the efficiency of clinical practice and health services [1]. This highlights the importance of novel machine learning techniques for EHR data, which can be integrated with mobile apps to provide personalized guidance for purposes ranging from early diagnosis to support for lifestyle change [2]. The latter is specifically relevant to reduce the cost of chronic diseases in the face of an aging population; for instance, the annual economic cost of diabetes in the U.S. is approximately $250 billion [3].

An important question is how to estimate a patient's response to a given treatment, comparing the patient's data from before and after the treatment. This is particularly challenging when the response is a continuous curve, for example a time-series of a biological marker. Such a response may be modeled using Gaussian processes [4] or neural networks [5]. The treatment may be a continuous dose function [6] or a discrete event [7], [8]. Treatment data are usually sparse, and hence it is essential to share relevant information in a probabilistic model. A latent trajectory model of [9] uses additive components to explain variation on population and individual levels. Conditional random fields can be incorporated to further capture correlations between different treatment types [10], and multivariate response curves can be modeled by learning latent structure [6] shared across the outcomes.

Despite much recent attention, there are still crucial issues in treatment-response estimation that have not been addressed. Most importantly, when the response is continuous, the treatments are consistently assumed known while in reality they may be perturbed by numerous factors. This problem dramatically escalates for user-reported data which potentially results in complete discredit of the findings [11]. The error is two-fold: first, there is measurement error in the treatment covariates; second, even the timings of the treatments might be known only approximately. Another issue arises from the complementary roles of the trend, i.e., the evolution of the outcome assuming no treatment, and the treatment response. When modeled and trained jointly, too flexible a trend may override the treatment response, and therefore in practice these two components are often trained separately.

To address the mentioned shortcomings, we introduce errors-in-variables (EIV) framework for modeling of continuous treatment-response trajectories. The EIV models account for measurement errors not only in the response, as common regression, but also in the inputs [12], [13]. They are closely related to latent-variable models in machine learning [14], modeling the

Manuscript received October 28, 2019; revised March 2, 2020; accepted April 6, 2020. Date of publication April 20, 2020; date of current version January 5, 2021. This work was supported by the BusinessFinland under Grant 884/31/2018) and the Academy of Finland under Grants 286607 and 294015). AJ was funded by the Academy of Finland under Grant 314457 and The Finnish Medical Foundation. KP was funded by the Academy of Finland under Grants 314383 and 266286 , the Academy of Finland Centre of Excellence in Research on Mitochondria, Metabolism and Disease (FinMIT; under Grant 272376), the Finnish Medical Foundation, the Gyllenberg Foundation, the Novo Nordisk Foundation under Grants NNF17OC0027232 and NNF10OC1013354), the Finnish Diabetes Research Foundation, the Finnish Foundation for Cardiovascular Research, the University of Helsinki and Helsinki University Hospital. *(Corresponding author: Pekka Marttinen.)*

Guangyi Zhang, Reza A. Ashrafi, and Pekka Marttinen are with the Department of Computer Science, Aalto University 02150, Espoo, Finland (e-mail: guangyi.zhang@aalto.fi; reza.ashrafi@aalto.fi; pekka.marttinen@aalto.fi).

Anne Juuti and Kirsi Pietiläinen are with the Medical School, University of Helsinki, 00100 Helsinki, Finland (e-mail: anne.juuti@hus.fi; kirsi.pietilainen@helsinki.fi).

This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this paper are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2020.2987323

unobserved true values underlying the noisy observations. Our contributions can be summarized as follows:

- We formulate an EIV model for personalized treatment-response trajectories, where a treatment comprises a vector of noisy covariates and treatment times are uncertain.
- We introduce an interpretable hierarchical prior on the treatment effects that efficiently shares information between individuals and allows joint training of the full model, appropriately balancing the trend and the responses.
- In a challenging application, representative of the current technological mega-trend of self-monitoring data from wearable devices, we show our method can meaningfully estimate the personalized impact of diet on continuous blood glucose measurements.

This paper is organized as follows. Next section reviews related work. In Section III, we discuss the components of the proposed model. Section IV presents results on both simulated and real-world data, and finally Section V concludes the paper. The code and the dataset used in the analyses are available at https://github.com/Guangyi-Zhang/eiv-treatment-response, which allow the full reproducibility of our results.

II. Related Work

Treatment response: Recurrent neural networks have been used for estimating treatment effects from time series data, e.g., [15], [16]. Besides machine learning, the problem of treatment response estimation has been studied in various fields, including informatics for medicine and social sciences, where the data-driven approach can bring advantages compared to experimental trials [17]. Individual-level treatment response prediction has been studied for schizophrenia [18], for Parkinson's disease from wearable sensor data [19], and depression [20]. An empirical comparison of classifiers for treatment-response prediction for chemoradiotherapy appears in [21]. Treatment response models have also been studies in social sciences, e.g., [22] and [23].

Mechanistic models: In contrast to the data-driven approaches used in machine learning, mechanistic models use substantial knowledge of a specific problem to characterize the system with differential equations, and inference is done for example using filtering algorithms. Similar to our application, [24] and [25] study blood glucose dynamics, affected by nutrition and other factors. Other examples are computational models for the physiological mechanisms of type-2 diabetes [26], and drug responses of cancer cell lines [27].

EIV models for treatment response: Estimating different types of regression functions under measument error has been studied recently [28], [29]. However, little work has focused on treatment-response estimation with EIV. Examples include inferring causal directions using EIV without conditioning on specific treatments [30] and predicting standardized test scores using student covariates [31]. Article [32] demonstrates the devastating impact of ingoring EIV for a binary response. An EIV model has also been used to quantify uncertainty when detecting treatment changes for liver metastases [33].

None of these works address the problem of estimating the impact of a multivariate vector of covariates on a continuous response with measurement error in covariates and uncertainty in treatment timing, the topic of this paper.

III. Methods

In this section, we first review EIV models on a general level. Then we describe the components of our model for personalized treatment-response trajectories: a hierarchical prior on parametric response functions, a Gaussian process for the trend, and measurement error models. We conclude the section by discussing the causal interpretation of the model. Throughout, we present the model in generic terms, but also outline the specific model used in Section IV-B to estimate the impact of diet on continuous blood glucose measurements.

Our model is fully Bayesian, yielding uncertainty estimates for all parameters, essential in scientific applications. Inference is done using Markov chain Monte Carlo (MCMC) with the state-of-the-art No-U-Turn (NUTS) sampler [34] implemented in software PyMC3 [35]. Implementation details are discussed below and in the Supplement, and can be inspected in full in the published code.

A. Errors-in-Variables Models

EIV models, a.k.a. measurement error models, are regression or classification models that, in contrast to most existing models, account for errors not only in the output variable but also in inputs [36], [37]. Though commonly neglected, input mismeasurement may be extremely harmful. For example in simple linear regression it leads to biased estimates that can not be corrected for even with an infinite sample, while, on the other hand, unbiased homoscedastic error in the output variable only induces additional variability [36]. A graphical model for a general EIV model is presented in Fig. 1(a), where X^* and Y^* represent the true values of the inputs and the output, and X and Y are the corresponding noisy observations. The most important type of mismeasurement is *classical error*, where it is assumed that the error term is independent of the true value.

Except for the simplest case of linear regression, EIV modeling almost always requires auxiliary information or data, e.g., instrumental variables or repeated measurements, to correct for the mismeasurement bias in estimation. However, without additional data, Bayesian EIV modeling is currently the most powerful and flexible approach, as it allows incorporating additional information in the form of distributional assumptions [37]. In this work, we adopt the Bayesian approach.

Mathematically, the measurement error mechanism is defined as the distribution of the noisy observed input, X, given the true unobserved input, X^*. The joint distribution of the model factorizes accordingly as:

$$P(X^*, Y^*, X, Y, \Theta) = P(X|X^*, \theta_M)P(Y|Y^*, \theta_N) \times P(Y^*|X^*, \theta_R)P(X^*|\theta_E)P(\Theta), \quad (1)$$

where $P(X|X^*, \theta_M)$ and $P(Y|Y^*, \theta_N)$ are called *error* or *measurement models*, $P(Y^*|X^*, \theta_R)$ is a *response* or *outcome*

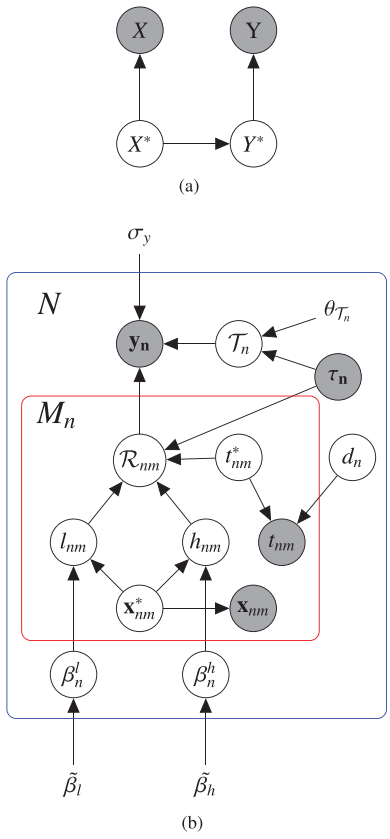


Fig. 1. (a) General formulation of the EIV model. For clarity, parameters associated with the distributions are not shown. (b) Proposed model for personalized treatment-response trajectories. Details of the model are discussed in Section III.

model, $P(X^*|\theta_E)$ is an *exposure model*, and $\Theta = (\theta_M, \theta_N, \theta_R, \theta_E)$ are the corresponding parameters. Bayes theorem can be used to infer the unknown parameters and unobserved true values of the variables.

$$P(X^*, Y^*, \Theta|X, Y) \propto P(\Theta) \prod_i^N P(X_i|X_i^*, \theta_M) P(Y_i|Y_i^*, \theta_N) P(Y_i^*|X_i^*, \theta_R) P(X_i^*, \theta_E). \tag{2}$$

If the exposure model is noninformative and the measurement model is symmetric, i.e., $P(X_i|X_i^*, \theta_M) = P(X_i^*|X_i, \theta_M)$, then the Bayesian modeling of classical error is equivalent to another class of mismeasurement techniques known as *Berkson error modeling* [36].

$$P(X^*, Y^*, \Theta|X, Y) \propto P(\Theta) \prod_i^N P(X_i^*|X_i, \theta_M) \times P(Y_i|Y_i^*, \theta_N) P(Y_i^*|X_i^*, \theta_R).$$

A well-known difficulty with EIV models is that they are often nonidentifiable [37], i.e. there are more than one set of values for the unknowns leading to the same model. This can be understood intuitively by noticing that the model stays the same if we multiply the linear regression coefficients by a constant factor and at the same time divide the estimated true values of inputs by the same factor. Therefore, to achieve identifiability, some crucial information about measurement model has to be assumed or estimated, e.g., the variance of a classical additive error in simple linear regression [36]. The Bayesian paradigm offers a unique solution to the nonidentifiability of the EIV models, as long as mismeasurement is modest and the prior is sufficiently good [38].

B. Model for Personalized Treatment-Response Trajectories

Notation: A graph of our model for treatment-response trajectories is presented in Fig. 1b. We assume there are N patients, and a trajectory consisting of a time series of length G_n of the outcome (e.g. blood glucose) is observed for each individual:

$$\mathbf{y_n} = (y_{n1}, \ldots, y_{nG_n})^T, n = 1, \ldots, N.$$

These measurements have been taken at times

$$\tau_\mathbf{n} = (\tau_{n1}, \ldots, \tau_{nG_n})^T, n = 1, \ldots, N.$$

Furthermore, each patient has M_n observed treatments (e.g. meals eaten), indexed by $m \in 1, \ldots, M_n$, where each treatment is characterized by P covariates:

$$\mathbf{x}_{nm} = (x_{nm1}, \ldots, x_{nmP})^T, \text{for all } m, n,$$

and the corresponding recorded treatment times are

$$\mathbf{t}_n = (t_{n1}, \ldots, t_{nM_n})^T, \text{for all } n.$$

Here, $\mathbf{x}_{nm}$ and t_{nm} are assumed to be noisy observations of the treatment covariates and timings, and their true unobserved values are denoted by $\mathbf{x}^*_{nm}$ and t^*_{nm}, respectively.

Outcome model: We model the observed outcome trajectory of individual n, $\mathbf{y}_n$, as

$$\mathbf{y}_n = \mathcal{T}_n + \sum_m \mathcal{R}_{nm} + \mathbf{e},$$

where $\mathcal{T}_n \in \mathbb{R}^{G_n}$ is a counterfactual trend (i.e. it describes the evolution of the outcome had the treatment not been taken), $\mathcal{R}_{nm} \in \mathbb{R}^{G_n}$ is the additive response to the mth treatment, and $\mathbf{e} = (e_1, \ldots, e_{G_n})^T$ is the vector of errors with $e_i \sim N(0, \sigma_y^2)$. We note that the sum of the trend and the responses can be viewed as a trajectory for a 'clean' outcome (omitted from Fig. 1b), of which a version $\mathbf{y_n}$ corrupted by Gaussian noise is observed.

Response function: Response functions specify how treatments affect the outcome over time, and they should be specified to suit the application at hand, balancing flexibility, interpretability, etc. For example, if interpretability is not needed and the amount of data is large, non-parametric functions that learn the shape of the response are attractive. On the other hand, parametric functions are suitable when data are scarce, and they are often interpretable, which is valuable in itself but also helps specifying prior knowledge to improve accuracy. In the application of learning the impact of meals on blood glucose (Section IV-B), we model the treatment response using a bell-shaped parametric function

$$\mathcal{R}_{nm} := f(\Delta_{nm}, h_{nm}, l_{nm}) := h_{nm} \exp\left(\frac{-0.5(\Delta_{nm} - 3l_{nm})^2}{l_{nm}^2}\right), \tag{3}$$

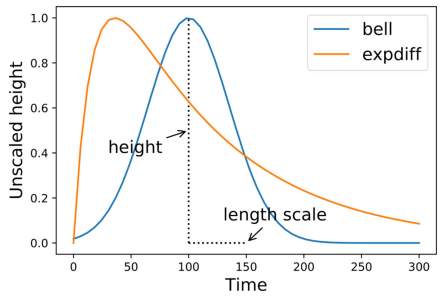


Fig. 2. Alternative response functions. The blue one is used in this work, while the orange one was used in [8]. The advantage of the blue response function is that it has only two parameters, both of which have intuitive interpretations.

where a lag vector $\Delta_{nm} = \boldsymbol{\tau}_n - t^*_{nm}$ represents the time since a specific treatment. The shape of this response is shown in Fig. 2 and it is determined by two parameters h_{nm} and l_{nm} with straightforward interpretations: h_{nm} is the height of the response, and l_{nm} is the length-scale which is proportional to the 'width' or 'duration' of the response. The main challenge in our application is sparsity and noisiness of data, with only 13 individuals and on average 10 meals per patient. We also tried a more flexible three-parameter response used in [8], which allows a skewed response (see Fig. 2), but this model suffered from convergence problems, for which reason we selected the simpler alternative.

In applications it is often of interest to measure how the response depends on treatment covariates, and therefore we allow these parameters to depend on the covariates:

$$\begin{aligned} h_{nm} &= (\beta^h_n)^T \mathbf{x}^*_{nm}, \text{ and} \\ l_{nm} &= (\beta^l_n)^T \mathbf{x}^*_{nm}, \text{ for all } n, m. \end{aligned} \tag{4}$$

In Equation (4), the coefficient vectors $\beta^h_n, \beta^l_n \in \mathbb{R}^P$ represent the *personalized impact* of each of the P covariates on the height or width of the response for the nth individual. To share information across individuals, we introduce a Bayesian hierarchical prior, see [39], and assume that the personalized height and length-scale coefficients, β^h_n and β^l_n, are drawn from common distributions:

$$\beta^h_n \sim N_P(\tilde{\beta}_h, \Sigma_h) \quad \text{and} \quad \beta^l_n \sim N_P(\tilde{\beta}_l, \Sigma_l).$$

A hyperprior is further placed on the mean parameters of these distributions:

$$\tilde{\beta}_h \sim N_P(\mathbf{0}, \tilde{\Sigma}_h) \quad \text{and} \quad \tilde{\beta}_l \sim N_P(\mathbf{0}, \tilde{\Sigma}_l)$$

The hierarchical prior introduces shrinkage and facilitates estimation of the personalized coefficients with limited data. Further details are given in the Supplementary material.

Counterfactual trend: A counterfactual trend represents the outcome assuming no treatment has been taken. It has to be sufficiently flexible to handle any variation in the outcome that is not accounted for by the treatments. In this paper, we model the trend $\mathcal{T}_n(t)$ for individual n using a Gaussian Process (GP) [4]:

$$\mathcal{T}_n(t) \sim \mathcal{GP}\left(0, k(t, t'|\theta_{\mathcal{T}_n})\right),$$

where $\theta_{\mathcal{T}_n}$ are parameters associated with the kernel function $k(x, x'|\theta_{\mathcal{T}_n})$. GPs are non-parametric regression models with well-known closed-form formulas for posterior estimation, which they inherit from the Normal distribution by assuming all training and test data follow a joint Normal distribution. For example, if

$$\mathcal{S}_n = \mathbf{y}_n - \sum_m \mathcal{R}_{nm}$$

is the residual of the outcome after subtracting the impact of the treatment responses, then

$$\begin{aligned} \mathcal{T}_n(t)|\mathcal{S}_n &\sim N(\mu_*, \Sigma_*), \quad \text{where} \\ \mu_* &= k(\tau_\mathbf{n}, t)^T K(\tau_\mathbf{n}, \tau_\mathbf{n})^{-1} \mathcal{S}_n, \quad \text{and} \\ \Sigma_* &= k(t, t) - k(\tau_\mathbf{n}, t)^T K(\tau_\mathbf{n}, \tau_\mathbf{n})^{-1} k(\tau_\mathbf{n}, t). \end{aligned}$$

We refer the reader to [4] for more details about GPs. As the kernel, we use the sum of Squared Exponential (SE) and constant kernels, where the former equips the GP with desired smoothness, and the latter enables meaningful extrapolation to regions where no input points have been observed. To speed up computation, we use a sparse GP [4] instead of a full GP, which samples a small set of inducing points uniformly from $\tau_\mathbf{n}$ to achieve a low-rank approximation of $K(\tau_\mathbf{n}, \tau_\mathbf{n})$ and its inverse.

When jointly training the treatment response functions and the trend, the trend must be properly regularized, not to explain away the treatments. We address this by using priors that discourage very short GP length-scales. This corresponds to the assumption that treatment effects dominate short-term variation in the response immediately after the treatment. If there are other strong causes of short-term variation, not represented by the treatments, then a more flexible GP trend could improve the short-term predictions, but prevent the estimation of the treatment effects. A detailed prior specification is provided in the Supplementary material.

Measurement models: Measurement models describe error in observations. With self-reported data both covariates and the timing of a treatment may be uncertain. To account for the uncertainty in treatment timing, we assume:

$$t_{nm} \sim N\left(t^*_{nm} + d_n, (\sigma^t_n)^2\right), \quad \text{for all } n, m.$$

In words, the observed time t_{nm} is obtained from the true time t^*_{nm} by shifting it with a bias term d_n, and adding Gaussian noise. The bias term d_n represents reporting habits of different individuals. For example, in the blood glucose application in Section IV-B, some individuals may systematically report their meal after eating, while others may do this before eating.

Different models are possible for treatment covariates, depending on the assumptions and data available [37]. Here we assume a simple perturbation on the *amount* of treatment:

$$\begin{aligned} \mathbf{x}_{nm} &= \mathbf{x}^*_{nm} \delta_{nm}, \quad \text{where} \\ \delta_{nm} &\sim LogNormal(0, \sigma^2_x), \quad \text{for all } n, m. \end{aligned} \tag{5}$$

The coefficient δ_{nm} represents the error in the mth treatment of the nth individual. Intuition in the blood glucose application is that users are able to report correctly what they have eaten, but not how much. While the model (5) captures our understanding of the type of mismeasurement expected in our data, more complicated models could also be justified, but they would

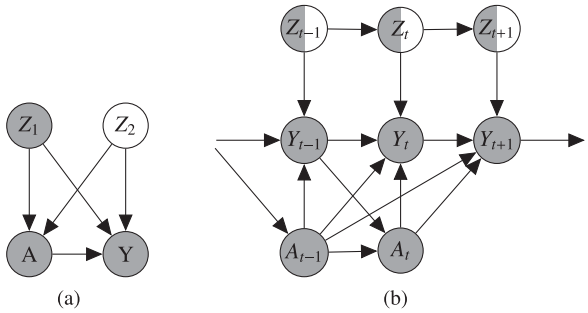


Fig. 3. a) Graphical model for a cross-sectional case, showing action (A), response (Y), and observed and hidden confounders (Z_1 and Z_2); b) over-time response with a single treatment, where confounders Z can be either observed or hidden.

require stronger additional assumptions to resolve the nonidentifiability of the EIV models. The model in (5) is identifiable and can be trained with relatively little data, as we demonstrate in Section IV. Further details, e.g., prior distributions for $\mathbf{x}^*_{nm}$, t^*_{nm}, and d_n, are provided in the Supplementary material.

C. Causal Assumptions and Interpretation

We briefly review results related to estimation of causal effects from observational data on treatment-response trajectories [8], [40], [41], to enable a user of our method to judge to what extent the effects found may or may not be interpreted causally. The causal effect of an action A (e.g. a treatment) on Y is defined as $P(Y = y|do(A = a))$, where the $do(\cdot)$ operator represents a manipulation of A to value a. The key assumption is that there are *no unmeasured confounders* (NUC), such as Z_2 in Fig. 3a. Without Z_2, the causal effect of A on Y can be estimated from observational data using the adjustment formula:

$$P(Y = y|do(A = a)) = \sum_{z_1} P(Y = y|A = a, Z_1 = z_1)P(Z_1 = z_1).$$

Time-varying treatments (Fig. 3b) further face an issue of treatment-confounder feedback [41], which means that hidden confounders do not have to affect A directly to create a spurious correlation between an action and future observations. A generalized adjustment formula, g-formula [41], can still be used to calculate $P(\bar{Y}_{\geq t}|do(A_{t-1}, A_t))$.

A useful result, applicable with our model, is to use the model to estimate $P(\bar{Y}_{\geq t}|\bar{A}_{\leq t}, \bar{Y}_{<t})$ from observational data. Then, assuming NUC, the following holds [41]:

$$P(\bar{Y}_{\geq t}|do(A_t), \bar{A}_{<t}, \bar{Y}_{<t}) = P(\bar{Y}_{\geq t}|\bar{A}_{\leq t}, \bar{Y}_{<t}). \quad (6)$$

In words, conditionally on the history of treatments and the outcome (and relevant observed confounders not shown in the formula), the causal impact of the most recent treatment on future outcomes can be estimated from observational data. This *short-term effect* [41] can be used, e.g., to select between alternative treatments available at a certain point in time, when the relevant history of the individual is known.

In Section IV-B we study the impact of diet on blood glucose. Based on domain knowledge, we know that diet is a prominent cause for changes in blood glucose. Furthermore, in our data we often see a rapid increase and decrease in glucose after a meal. Therefore, it is plausible to assume that meals affect blood glucose causally. In general, causal assumptions can not be verified from observational data, and it is possible that some confounder affects both glucose and diet. The effect of any such confounder is expected to be small compared to diet. Hence, causal interpretation of our results seems reasonable, but assertions of this can not be made. With modern wearable self-monitoring devices it will be possible to measure all relevant factors that could affect blood glucose much more comprehensively, and the NUC assumption is reasonable. Our model is straightforward to extend to such data.

IV. Results

In this section, we first validate our model using simulated data, and then apply it to a real-world dataset comprising diet and continuous blood glucose measurements. Throughout, we compare four models, in an increasing order of complexity (later models include the previous as special cases):

- $\mathcal{M}_{ind}$: Separate models for individuals.
- $\mathcal{M}_{hier}$: Model with the hierarchical prior for the responses to share information across individuals.
- $\mathcal{M}_{hier+time}$: Time uncertainty included.
- $\mathcal{M}_{hier+time+cov}$: Uncertainty in covariates included.

A. Validation and Identifiability of the Models With Simulated Data

As the first simple experiment we study the identifiability of the EIV model when there is measurement error in covariates. We simulate artificial data using a toy model specified as the sum of a linear trend and the parametric treatment response from Equation (3). The dimension of treatment covariates is here set to 2, and each input is perturbed with an additive term drawn from $N(1, 0.2^2)$. We analyze the data using the EIV model that assumes measurement error, and a model that disregards the noise in the covariates. Results and details for this simple setup are presented in the Supplementary material, and they show that the EIV model recovers all true inputs and effect sizes with high accuracy, while the model that neglects the noise leads to biased coefficient estimates and wide confidence intervals.

To study the accuracy and identifiability of our method in a more realistic simulated setup, we first fit our model to the real-world data from Subsection IV-B, and use the fitted model to simulate additional data for two individuals. We perturb half of the inputs according to Equation (5) and let the model use the perturbed inputs and try to recover the true inputs and parameters. The performance of all models depends on the relative contributions of the trend and responses, and we scale up the response with a factor of 5, which facilitates a meaningful comparison.

Results for one individual are shown in Fig. 4. Results for the other individual, and replicated results for both, initialized with a different seed to assess variability in training, are shown in the Supplement. We see that the direction of each non-zero perturbation is estimated correctly (left panel), also for the other individual (Supplement). However, if there is no perturbation, non-zero perturbations may still be learned, introducing noise.

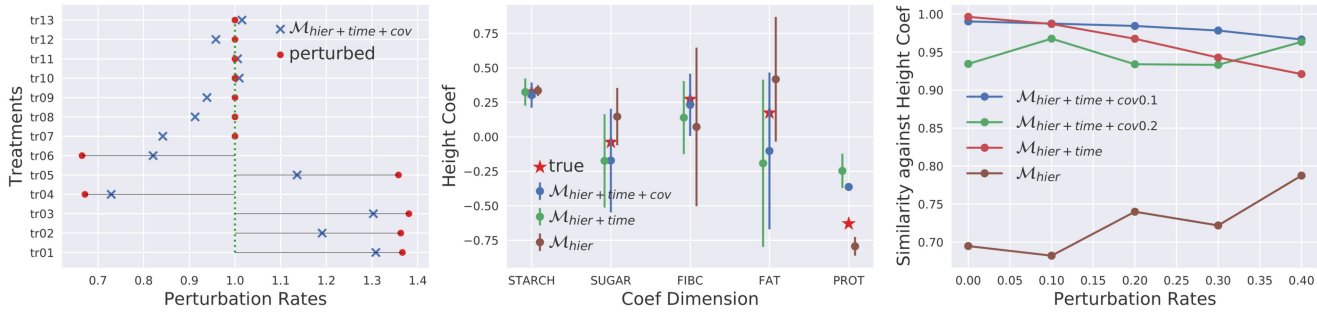


Fig. 4. Simulation results. *Left:* true and estimated perturbations for one individual (the other one shown in Supplement); *Center:* true and estimated coefficients (mean +/− SD) for the height of the response for the 5 covariates; and *Right:* Cosine similarity of concatenated height coefficient vectors from all individuals against the true value with different levels of perturbation (larger value is better). Two different prior SDs, 0.1 and 0.2, were considered for model $\mathcal{M}_{hier+time+cov}$.

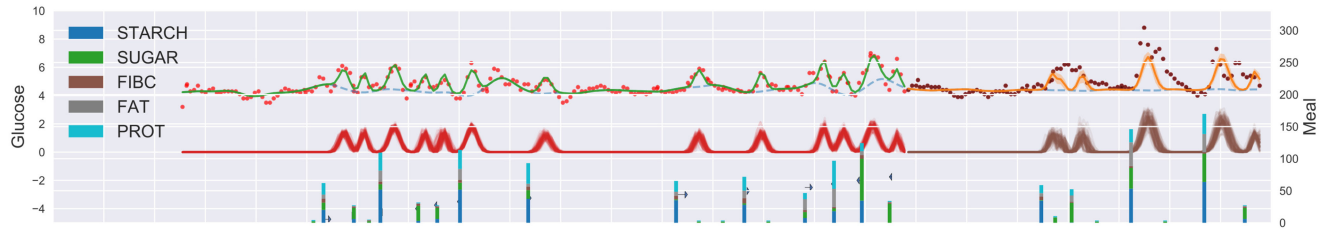


Fig. 5. Visualization of the 3-day time series for one patient. Red and brown dots represent glucose markers in the training and test sets, respectively. Meals are indicated by vertical bars, colored according to amounts of different nutrients in the meal. The green curve is the final fitted trajectory, and it is a combination of the dashed blue line, a counterfactual trend, and the mean of red lines, which are posterior samples of treatment responses. Horizontal arrows associated with the meals show the estimated difference between true and observed meal times.

This trade-off between flexibility and overfitting highlights the importance of carefully validating the model to suit the amount and complexity of data. The regression coefficients are estimated accurately by the EIV model (center), and the benefit from using EIV increases when the size of the perturbation increases (right). However, a too loose EIV prior (large SD) may actually harm the performance by introducing noise, when the true perturbation is small.

B. Application to Continuous Blood Glucose Trajectories

The data set contains blood glucose measurements and dietary records for 13 non-diabetic individuals across three days. The anonymized data were provided by the Obesity Research Unit at the University of Helsinki. The study was approved by the Ethics Committee of Helsinki and Uusimaa Hospital District (HUS/1706/2016) and by Helsinki University Hospital research review board (HUS269/2017). The real-valued blood glucose measurements were collected by a portable continuous glucose monitoring system at approximately every fifteen minutes. The dietary records consist of contents and times of all meals eaten during the 3-day study period. Each meal has been processed into amounts of five nutrients: starch, sugar, fiber, fat, and protein. Our goal is to learn how these nutrients influence blood glucose. Both meal times and the amounts of food eaten are uncertain, as they are reported by the users, which motivates the use of EIV models. The data (and results) for one individual are visualized in Fig. 5, and for all other individuals in the Supplement. Some markers may be missing due to device errors or when a user has removed the device.

Metrics: The models are trained using data from the first two days, and the third day is used for testing. The performance of treatment-response estimation is quantified using five metrics $\mathrm{M}_i, i \in \{1, \ldots, 5\}$. The first two metrics quantify the relative contributions of the trend and treatment responses in the model fitted to the training data. In detail, M_1 is the proportion of variance explained by the trend:

$$\mathrm{M}_1 = \frac{1}{N} \sum_n \frac{Var(\mathcal{T}_n)}{Var(\mathbf{y}_n)}.$$

M_2 indicates how much more is explained when also the treatment responses are included:

$$\mathrm{M}_2 = \frac{1}{N} \sum_n \frac{Var\left(\mathcal{T}_n + \sum_m \mathcal{R}_{nm}\right)}{Var(\mathbf{y}_n)} - \mathrm{M}_1.$$

A large M_1 means that the outcome is mostly explained by the trend, and a small M_2 represents an inactive response.

Metrics M_3 and M_4 are the mean squared errors in the training and test data. They are calculated for all patients for whom M_2 indicates that the response has been properly learned. Thus one patient, with $\mathrm{M}_2 \approx 0.05$ for the baseline model $\mathcal{M}_{hier}$ is excluded from MSE calculations (other patients have $\mathrm{M}_2 > 0.3$).

Because M_4 measures pointwise error, it may be misleadingly high even when the shape of the response is estimated perfectly, if its location is inaccurate. Metric M_5 is insensitive to this inaccuracy of location, and it measures the absolute error in variance between predicted response and outcome:

$$\mathrm{M}_5 = \frac{1}{N} \sum_n \left| Var\left(\sum_m \mathcal{R}_{nm}\right) - Var(\mathbf{y}_n) \right|$$

TABLE I

COMPARISON OF MODELS USING THE REAL-WORLD GLUCOSE DATA. METRICS M_1 TO M_5 ARE DEFINED IN THE TEXT. P-VALUE TESTS IF OTHER MODELS ARE BETTER THAN $\mathcal{M}_{hier}$ IN TERMS OF M_4. PVE: PROPORTION OF VARIANCE EXPLAINED, LOO: LEAVE-ONE-OUT CROSS-VALIDATION, PLOO: THE ESTIMATED EFFECTIVE NUMBER OF PARAMETERS, SE-LOO: THE STANDARD ERROR IN THE LOO COMPUTATIONS

	M_1 PVE Trend	M_2 PVE Resp	M_3 MSE Train	M_4 MSE Test	M_5 ΔVar Test	p-value U-test	LOO	pLOO	SE LOO
$\mathcal{M}_{ind}$	0.361	0.342	0.149	1.695	0.927	1.00	3549.64	246.64	318.8
$\mathcal{M}_{hier}$	0.359	0.339	0.159	0.752	0.391	-	3587.87	214.62	317.28
$\mathcal{M}_{hier+time}$	0.350	0.424	0.098	**0.738**	0.377	**3.24e-4**	**2869.91**	342.24	265.09
$\mathcal{M}_{hier+time+cov}$	**0.344**	**0.428**	**0.098**	0.743	**0.366**	4.66e-3	2994.98	465.47	333.7

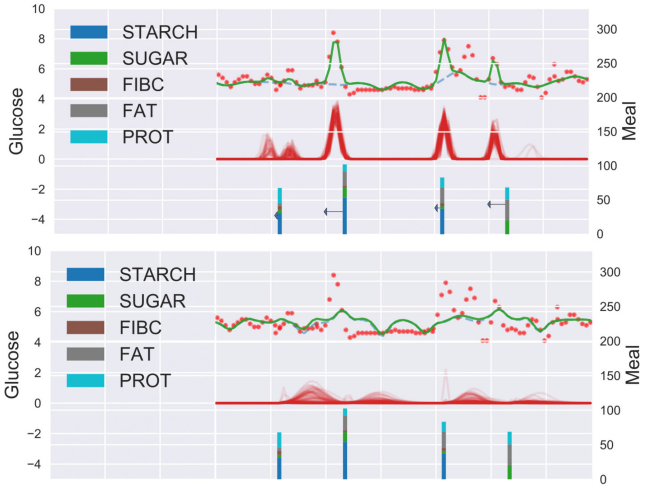
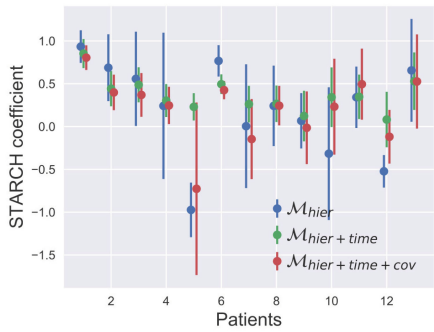


Fig. 6. Demonstration of time uncertainty modeling for one individual for one day. *Top:* Results using $\mathcal{M}_{hier+time}$, where arrows indicate the estimated difference between the true and observed meal times; *Bottom:* Results using $\mathcal{M}_{hier}$.

	Mean	Std
STARCH	9.57	4.29
SUGAR	6.36	4.97
FIBC	2.83	4.95
FAT	5.29	5.49
PROT	−10.13	3.90

(a)



(b)

Fig. 7. *a*) Impact of different nutrients on the area under the response curve, ΔA_{np}, averaged across the patients; *b*) Posterior uncertainty (mean +/- SD) of the personalized coefficients representing the impact of starch on the height of the response.

Because our interest is in estimation of the treatment response, and not the trend, we calculate all metrics in windows spanning from one hour before to three hours after each meal.

We use the Mann–Whitney U-test [42] to test if other models are better than $\mathcal{M}_{hier}$ in terms of test error M_4. The reason for using $\mathcal{M}_{hier}$ as the baseline is our argument that EIV modeling is beneficial when estimating treatment-response trajectories, and $\mathcal{M}_{hier}$ is otherwise the same as $\mathcal{M}_{hier+time}$ and $\mathcal{M}_{hier+time+cov}$ except that it the does not include the EIV components. We also compare the models using the state-of-the-art information criterion for predictive accuracy, leave-one-out cross-validation (LOO) [43].

Results: Result are shown in Table I. We see that all models outperform the non-hierarchical baseline $\mathcal{M}_{ind}$ by a large margin. Furthermore, taking treatment time inaccuracy into account in $\mathcal{M}_{hier+time}$ improves significantly over the non-EIV model $\mathcal{M}_{hier}$. In fact, estimation of the response fails completely for some individuals without time EIV; the results with and without time uncertainty modeling for one such case are shown in Fig. 6. On the other hand, taking uncertainty in covariates into account does not notably improve accuracy, which is likely caused by a combination of increased flexibility and limited amount of data. Overall, models with EIV component outperform the model without EIV in all metrics.

Interpretability of personalized treatment response is also of great interest; for instance, understanding how an individual's glucose level changes if she eats one more unit of sugar. The overall goal of glucose monitoring is to keep the glucose level in a given range, and both the amount of excess and the duration of the hyperglycemic state are clinically important. Hence, a sensible parameter to consider is the impact of different nutrients on the *area* of the response curve. Though this is not a parameter of our model, it is straightforward to derive the personalized increase in response area due to one unit increase of a specific nutrient ΔA_{np}, $n = 1, \ldots, N$, $p = 1, \ldots, P$, using the estimated coefficients for height and width, which are modeled explicitly (see Supplement for a derivation). Overall, starch and sugar have the strongest positive impact on glucose (Fig. 7a), consistent with the understanding that carbohydrates increase blood glucose [44]. Protein, on the other hand, has a negative impact, which has been observed before and might represent a complex short-term interaction between nutrients [45]. An advantage of our model is that we get *personalized* coefficients for each individual together with their associated uncertainties, as shown for starch in Fig. 7b, and for the other nutrients in the Supplement. Importantly, models with EIV have much narrower confidence intervals for the estimated effects, meaning that they are estimated more accurately, thanks to the increased flexibility that allows fitting the complex data.

V. Conclusion

We presented a hierarchical EIV model to estimate personalized treatment-response trajectories when the covariates and timings of the treatments were imprecise. Our model had superior performance in simulated and real-world data on various metrics, and yielded interpretable and meaningful estimates of the personalized effects of treatment covariates, valuable in applications. Future directions include studying the identifiability of EIV models in the context of continuous treatment-response trajectories further and extending the model to include interactions between covariates and other unmeasured confounders, such as physical activity, for causal completeness.

References

[1] J. Powell and I. Buchan, "Electronic health records should support clinical research," *J. Med. Internet Res.*, vol. 7, no. 1, pp. 93–118, 2005.

[2] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, and R. Ranganath, "Opportunities in machine learning for healthcare," 2018, *arXiv:1806.00388*.

[3] American Diabetes Association, "Economic costs of diabetes in the U.S. in 2012," *Diabetes Care*, vol. 36, no. 4, pp. 1033–1046, 2013.

[4] C. E. Rasmussen, "Gaussian processes in machine learning," in *Advanced Lectures on Machine Learning*. Berlin, Germany: Springer, 2004, pp. 63–71.

[5] B. Lim, "Forecasting treatment responses over time using recurrent marginal structural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2018, pp. 7494–7504.

[6] H. Soleimani, A. Subbaswamy, and S. Saria, "Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions," in *Proc. 33rd UAI Conf. Uncertainty Artif. Intell.*, 2017.

[7] Y. Xu, Y. Xu, and S. Saria, "A non-parametric bayesian approach for estimating treatment-response curves from sparse time series," in *Proc. Mach. Learn. Healthcare Conf.*, vol. 56, 2016, pp. 282–300.

[8] P. Schulam and S. Saria, "Reliable decision support using counterfactual models," in *Proc. Advances Neural Inf. Process. Syst.*, 2017, pp. 1697–1708.

[9] P. Schulam and S. Saria, "A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure," in *Proc. Advances Neural Inf. Process. Syst.*, 2015, pp. 748–756.

[10] P. Schulam and S. Saria, "Integrative analysis using coupled latent variable models for individualizing prognoses," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 8244–8278, 2016.

[11] B. Kreider, "Regression coefficient identification decay in the presence of infrequent classification errors," *Rev. Econ. Statist.*, vol. 92, no. 4, pp. 1017–1023, 2010.

[12] Z. Griliches, "Errors in variables and other unobservables," *Econometrica*, vol. 42, no. 6, pp. 971–998, 1974.

[13] Z. Griliches and J. A. Hausman, "Errors in variables in panel data," *J. Econometrics*, vol. 31, no. 1, pp. 93–118, 1986.

[14] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, 1st ed. Cambridge, MA, USA: MIT Press, 2013.

[15] I. Bica, A. M. Alaa, J. B. Jordon, and M. van der Schaar, "Estimating counterfactual treatment outcomes over time through adversarially balanced representations," in *Proc. 8th Int. Conf. Learn. Representations*, 2020.

[16] Y. Zhang, A. Bellot, and M. van der Schaar, "Learning overlapping representations for the estimation of individualized treatment effects," 2020, *arXiv:2001.04754*.

[17] I. C. Passos and B. Mwangi, "Machine learning-guided intervention trials to predict treatment response at an individual patient level: An important second step following randomized clinical trials," *Mol. Psychiatry*, vol. 25, pp. 701–702, 2020.

[18] B. Cao *et al.*, "Treatment response prediction and individualized identification of first-episode drug-naïve schizophrenia using brain functional connectivity," *Mol. Psychiatry*, vol. 25, pp. 906–913, 2020.

[19] I. Thomas *et al.*, "A treatment-response index from wearable sensors for quantifying parkinson's disease motor states," *IEEE J. Biomed. Health Informat.*, vol. 22, no. 5, pp. 1341–1349, Sep. 2018.

[20] R. Pearson, D. Pisner, B. Meyer, J. Shumake, and C. G. Beevers, "A machine learning ensemble to predict treatment outcomes following an internet intervention for depression," *Psychological Medicine*, vol. 49, pp. 2330–2341, 2019.

[21] T. M. Deist *et al.*, "Machine learning algorithms for outcome prediction in (chemo) radiotherapy: An empirical comparison of classifiers," *Med. Phys.*, vol. 45, no. 7, pp. 3449–3459, 2018.

[22] D. Card, "The causal effect of education on earnings," in *Handbook of Labor Economics*, vol. 3, 1st ed. New York, NY, USA: Elsevier, 1999, Part A, ch. 30, pp. 1801–1863.

[23] C. F. Manski, "Identification of treatment response with social interactions," *Econometrics J.*, vol. 16, no. 1, pp. S1–S23, 2013.

[24] N. P. Balakrishnan, L. Samavedham, and G. P. Rangaiah, "Personalized mechanistic models for exercise, meal and insulin interventions in children and adolescents with type 1 diabetes," *J. Theor. Biol.*, vol. 357, pp. 62–73, 2014.

[25] D. J. Albers, M. Levine, B. Gluckman, H. Ginsberg, G. Hripcsak, and L. Mamykina, "Personalized glucose forecasting for type 2 diabetes using data assimilation," *PLoS Comput. Biol.*, vol. 13, no. 4, 2017, Art. no. e1005232.

[26] J. Sarkar *et al.*, "A long-term mechanistic computational model of physiological factors driving the onset of type 2 diabetes in an individual," *PLOS ONE*, vol. 13, no. 2, pp. 1–37, 2018.

[27] F. Frohlich *et al.*, "Efficient parameter estimation enables the prediction of drug response using a mechanistic pan-cancer pathway model," *Cell Syst.*, vol. 7, no. 6, pp. 567–579, 2018.

[28] Y. Kang, X. Gong, J. Gao, and P. Qiu, "Errors-in-variables jump regression using local clustering," *Statist. Medicine*, vol. 38, no. 19, pp. 3642–3655, 2019.

[29] S. Zhou, D. Pati, T. Wang, Y. Yang, and R. J. Carroll, "Gaussian processes with errors in variables: Theory and computation," 2019, *arXiv:1910.06235*.

[30] Y. Zhang and G. Luo, "Inferring causal directions in errors-in-variables models," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 3152–3153.

[31] J. R. Lockwood and D. F. McCaffrey, "Correcting for test score measurement error in ANCOVA models for estimating treatment effects," *J. Educational Behavioral Statist.*, vol. 39, no. 1, pp. 22–52, 2014.

[32] D. Millimet, "The elephant in the corner: A cautionary tale about measurement error in treatment effects models," Institute for the Study of Labor, IZA, Bonn, Germany, Discussion Papers 5140, 2010.

[33] R. Pathak *et al.*, "A data-driven statistical model that estimates measurement uncertainty improves interpretation of ADC reproducibility: A multi-site study of liver metastases," *Scientific Rep.*, vol. 7, no. 1, pp. 14 084–14 094, 2017.

[34] M. D. Hoffman and A. Gelman, "The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1593–1623, 2014.

[35] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in Python using PyMC3," *Peer J. Comput. Sci.*, vol. 2, p. e55, 2016.

[36] R. J. Carroll, D. Ruppert, C. M. Crainiceanu, and L. A. Stefanski, *Measurement Error in Nonlinear Models: A Modern Perspective*. London, U.K.: Chapman & Hall, 2006.

[37] P. Gustafson, *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*, 1st ed. New York, NY, USA: CRC Press, 2004.

[38] P. Gustafson, N. D. Le, and R. Saskin, "Case-control analysis with partial knowledge of exposure misclassification probabilities," *Biometrics*, vol. 57, no. 2, pp. 598–609, 2001.

[39] A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*. London, U.K.: Chapman & Hall, 2013.

[40] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[41] J. M. R. Miguel and A. Hernán, "Causal inference," 2018, [Online]. Available: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/s

[42] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, pp. 50–60, 1947.

[43] A. Vehtari, A. Gelman, and J. Gabry, "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC," *Statist. Comput.*, vol. 27, no. 5, pp. 1413–1432, 2017.

[44] T. M. Wolever and J. B. Miller, "Sugars and blood glucose control," *Amer. J. Clin. Nutrition*, vol. 62, no. 1, pp. 212S–221S, 1995.

[45] A. Karamanlis *et al.*, "Effects of protein on glycemic and incretin responses and gastric emptying after oral glucose in healthy subjects," *Amer. J. Clin. Nutrition*, vol. 86, no. 5, pp. 1364–1368, 2007.